学校编码：10384　　分类号_____密级_____
学　　号：200331046　　UDC _____

厦门大学

硕 士 学 位 论 文

基于网格密度和空间划分树的聚类算法研究
The Study of Clustering Algorithm based on Grid-Density and Spatial Partition Tree

曾东海

指导教师姓名：米 红　教 授
专 业 名 称：模式识别与智能系统
论文提交日期：2006 年 4 月
论文答辩时间：2006 年 5 月
学位授予日期：2006 年　月

答辩委员会主席：_____

评　阅　人：_____

2006 年 4 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

学校编码：10384　　分类号_____密级_____
学　　号：200331046　　UDC _____

厦门大学

硕 士 学 位 论 文

基于网格密度和空间划分树的聚类算法研究
The Study of Clustering Algorithm based on Grid-Density and Spatial Partition Tree

曾东海

摘 要

在数据挖掘领域中，聚类分析是一项重要的研究课题。它既可以作为一个单独的工具用以发现数据库中数据分布的深层信息，也可以作为其他数据挖掘分析算法的一个预处理步骤，因此研究如何提高聚类算法的性能具有重要的意义。

本文在分析现有聚类算法特别是基于密度的聚类算法优缺点的基础上，结合空间索引技术，提出了一种新的基于格网密度和空间划分树的聚类算法（CGDSPT）；在聚类实验系统上，通过对多个样本数据集的实验结果的分析和算法的实际应用，验证了 CGDSPT 算法的有效性。本文的主要工作包括：

1、将现有聚类方法按照五大类进行了系统的评述，并对基于密度的几种经典算法做了详细的介绍。

2、通过对空间索引结构的综述，结合空间划分的特性，提出了一种基于空间划分的索引结构 SP-Tree。SP-Tree 有效地保存了数据的空间位置信息，为空间区域的邻域查询提供了极大的方便；同时它只索引非空单元格，不仅节省了存储空间还降低了算法的时间复杂性。

3、结合基于格网密度聚类算法的特性和空间索引的优点，文章提出一种基于格网密度和空间划分树的聚类算法。算法充分借助了网格和空间索引的优势，使算法的时间复杂度与数据规模近似呈现线性关系。同时该算法具有能发现任意形状的簇、对噪声数据和数据输入顺序不敏感等优良特性。

4、针对算法的参数设置问题，本文提出了一种根据样本数据的统计特性自行调整参数的方法，能有效地降低参数设置的难度，获得了较好的聚类效果。

5、针对聚类有效性评价问题，本文提出了一种基于簇密度的适合任意形状簇的聚类有效性指数，实验表明其能有效地指导用户调整参数以获得满意结果。

6、建立了一个聚类实验系统。在此系统上，利用多个样本集对本文提出的聚类算法进行详细的性能分析；将算法应用到中国分区域人口多维综合死亡模式的聚类中，并对聚类结果的区域性等特征进行了详尽分析。

关键词：聚类；网格密度；空间划分树

Abstract

Clustering analysis is an important research problem in the domain of data mining. It can be used not only as a separate technique to discover the information about data distribution, but also as the preprocessing of other data mining operations, therefore it is very meaningful to research how to boost the performance of clustering algorithms.

This thesis mainly studies a new clustering algorithm based on the grid-density and the spatial partition tree (CGDSPT) through analyzing many presented representative clustering algorithms especially the density-based clustering algorithm. We design and realize a clustering experimental system (MODE-CES) with the C# development tool. It is proved that the CGDSPT is efficient by analyzing experiments of many data sets. The primary research include as follows:

1. The presented clustering algorithms are divided to five classes and discussed systemically. And some density-based clustering algorithms are described in detail.

2. The spatial indexes are described and a novel spatial index structure (SP-Tree) is presented based on the spatial partition. The SP-Tree can keep the spatial location of the data efficiently that makes the region neighborhood search become facilitative. Meanwhile it only indexes the non-empty cells in the partitioned space that saves the memory and boosts the performance.

3. A clustering algorithm based on the grid-density and spatial partition tree (CGDSPT) is presented by assimilates the advantages of the based-density and based-grid clustering algorithm and the spatial index structure. CGDSPT is a high performance clustering algorithm whose computational complexity is linear-time. Meanwhile this algorithm have many others outstanding characteristics such as it is robust to outliers, can identify clusters having any shapes and wide variances in size , non-sensitive to the sequence of the sample.

4. Aiming at the issue of set the parameters correctly, we offer a novel method, which makes the parameters be changed with the statistic characteristic of the dataset,

to set the parameters. It can reduce the difficulty of the setting parameters and get perfect result.

5. Aiming at the clustering validity problem, we present a new clustering validity index based on the density of the cluster, which is fit for any shape cluster. And the experiments indicate that this index can help the user adjust the parameters in order to gain satisfied clustering result.

6. A clustering experimental system is built. And to evaluate the performance and effectivity of the algorithm proposed in this thesis, extensive of experiments have been done. Meanwhile, we apply this algorithm to the clustering of China district death pattern, and analyze the regionality character of the clustering-result and so on. The results show that our algorithm is a high performance and efficient clustering algorithm.

Key Words: Clustering; Grid-Density; Spatial Partition Tree

目 录

Contents

第一章 绪论

1.1 本文的选题背景及研究意义

20 世纪 90 年代以来，随着信息技术和数据库技术的迅猛发展，人们可以非常方便地获取和存储大量的数据。面对大规模的海量的数据，传统的数据分析工具（如管理信息系统）只能进行一些表层的处理（如查询、统计等），而不能获得数据之间的内在关系和隐含的信息。为了摆脱“数据丰富，知识贫乏”的困境，人们迫切需要一种能够智能地自动地把数据转换成有用信息和知识的技术和工具，这种对强有力数据分析工具的迫切需求使得数据挖掘技术应运而生。

聚类作为数据挖掘技术的主要方法之一，是一种重要的数据分析技术，它搜索并识别一个有限的种类集合或簇集合，从而描述数据。聚类分析与分类有所不同，聚类的目标是在没有任何先验知识的前提下，根据数据的相似性将数据聚合成不同的簇(或类)，使得相同簇中的元素尽可能相似，不同簇中的元素差别尽可能大，因此又被称为非监督分类(unsupervised classification)。其已经受到了越来越多关注。原因在于：

（1）在很多实际应用中，由于缺少形成模式类过程的知识，或者由于实际工作中的困难，收集并标志大型样本集是很费时费力的工作，所以我们也往往只能用没有类别标签的样本集进行工作。另一方面，存在很多应用，待分类模式的性质会随着时间发生缓慢的变化。如果这种性质的变化能在无监督的情况下捕捉到，分类器的性能就会大幅度提升。

（2）在一个广义的数据挖掘过程中，聚类分析往往是被作为最初的步骤，用于对数据分布和聚合特性的初步了解，以在此基础上进行其他数据挖掘操作。

聚类分析作为统计学的一个分支，已经被广泛研究了许多年。而且，聚类分析也已经广泛地应用到诸多领域中，包括模式识别、数据分析、图像处理以及市场研究[1]。通过聚类，人们能够识别密集的和稀疏的区域，因而发现全局的分布模式，以及数据属性之间的有趣的相互关系。在商务上，聚类能帮助市场分析人员从客户基本信息库中发现不同的客户群，并且用购买模式来刻画不同的客户群的特征。在生物学上，聚类能用于推导植物和动物的分类，对基因进行分类，获

得对种群中固有结构的认识。聚类在地球观测数据库中相似地区的确定，汽车保险单持有者的分组，及根据房屋的类型、价值和地理位置对一个城市中房屋的分组上也可以发挥作用。

聚类分析作为数据挖掘系统中的一个模块，既可以作为一个单独的工具以发现数据库中数据分布的深层信息，也可以作为其他数据挖掘分析算法的一个预处理步骤。

因此聚类分析已成为数据挖掘领域中的一个非常活跃的研究课题，虽然经过几十年的发展，它已经较为成熟，但仍还有许多值得研究发展的地方。鉴于以上的认识，在厦门大学“985”二期重点项目“智能化国防安全信息技术科技创新平台”的资助下，本文对基于格网密度的聚类算法进行了相关研究，为构建高效的聚类算法做出自己的努力。

1.2 什么是聚类

聚类(clustering)[2]就是将数据对象分组成为多个簇(cluster)，使得同一个簇中的对象之间具有较高的相似性(similarity)，而不同簇中的对象具有较大的相异性(dissimilarity)。一个好的聚类方法应产生具有如下特性的聚类结果：簇内的对象高度相似(high intra-class similarity)，而簇间的对象很少相似(low inter-class similarity)。

定义 1.1 假定一个数据对象由 d 个属性(也称为度量或变量)描述，则若干个具有 d 个属性的数据对象就构成了 d 维数据空间。在 d 维空间中，数据对象被称作 d 维数据点，则 d 维数据点 x 可表示为 $x=(x_1,x_2,\mathrm{L}\ ,x_d)$，其中 x_i 表示第 i 个属性值，d 表示空间的维数(dimensionality)。

定义 1.2 由 n 个 d 维数据点组成的集合(又称为 d 维数据集)S 可表示为 $S=(s_1,s_2,\mathrm{L}\ ,s_n)$，其中 $s_i=(s_{i1},s_{i2},\mathrm{L}\ ,s_{id})$，且 s_{ij} 表示第 i 个数据点的第 j 个属性值。

定义 1. 3 根据数据点之间的相似性，将 d 维数据集 V 划分成 $\{C_1,C_2,\mathrm{L}\ ,C_k\}$ 的过程称为聚类分析，其中 $k\le n, C_i\ne\phi, C_i\subseteq V$ (i=1, 2,，……k)，并且 $\bigcup_{i=1}^{k}C_i=V$ 。

这里，C_i 一般被称做类或簇。

聚类分析以相似性(或相异性)为基础来划分簇，但是数据对象之间的相似性(或相异性)没有唯一的定义。相似性(或相异性)的具体定义依赖于数据类型以及评价相似性(或相异性)的角度。

评价相似性的角度[3]有三种，分别是基于距离的((distance-based)、基于密度的(density-based)和基于连接的(linkage-based)。前两类通常适用于欧几里德空间(Euclidean space)，第三类则适用于任意度量空间(metric space)。

定义 1. 4 根据数据点之间的距离评价相似性：距离越短，相似性越大；反之，距离越长，相似性越小。

数据点 v_i 和 v_j 的距离 d_{ij}，必须满足以下条件：

（1） $d_{ij} \geq 0$，当且仅当 i=j 时等号才成立（非负性）；

（2） $d_{ij} = d_{ji}$（对称性）；

（3） $d_{ik} \leq d_{ij} + d_{jk}$，其中 $v_i \neq v_j \neq v_k$（三角不等性）。

满足上述条件的 d_{ij} 的取值在[0，$+\infty$)。d_{ij} 越小，v_i 和 v_j 的相似性越大；反之，d_{ij} 越大，v_i 和 v_j 的相似性越小。

聚类分析中常用的距离定义有：

①明氏（Minkowski）距离 $d_{ij} = \left(\sum_{k=1}^{d} \left| x_{ik} - x_{jk} \right|^{q} \right)^{1/q}$；

②曼哈顿（Manhattan）距离 $d_{ij} = \sum_{k=1}^{d} \left| x_{ik} - x_{jk} \right|$；

③欧氏（Euclidean）距离 $d_{ij} = \left(\sum_{k=1}^{d} \left| x_{ik} - x_{jk} \right|^{2} \right)^{1/2}$；

其中 q>0，d 是空间的维数。还有其他距离定义，在此不一一列举。

定义 1. 5 根据数据点的分布密度来评价相似性：当数据点属于相连的密集区域时，它们是相似的，可以划归至同一个簇；否则，它们不相似，因而不属于同一个簇。

计算密度的方法又分为基于最近邻的(Nearest-Neighbor, NN)方法和基于单

元的(cell-based)方法。

NN 方法根据数据集的自身分布来定义密度。它规定密集区域必须符合以下条件：以该密集区域中的任意一点为中心，某种长度为半径的球形区域中的点数必须大于等于某个阀值。

单元方法根据散落在单元中的点数来定义密度。它规定将数据空间划分若干个单元，容纳点数大于等于某个阀值的单元是密集单元，属于密集区域。

定义 1. 6 根据图(或超图)模型中边的连接程度来评价相似性：将数据集映射成图或超图，图中边(或超边)的连接程度反映了该条边(或超边)连接的顶点的相似程度，强连接的边(或超边)经过的数据点之间的相似性较大，因此可被划分至同一个簇。

依据评价相似度的角度，聚类算法也可相应地分为三类：认为簇是由距离靠近的点组成的基于距离的聚类算法；认为簇是由相连的密集区域组成的基于密度的聚类算法；认为簇是由强连接的边上的点组成的基于连接的聚类算法。

簇有多种表示方法[3]，比如代表点、核心点、单元等。基于距离的聚类算法大多采用代表点表示簇。可以只使用一个点来表示一个簇，这个点可以是该簇的质心(簇中所有点的平均值)，也可以是该簇中位置最接近中心的点，因此只能识别球形簇；也可以使用多个点来表示一个簇，这些代表点比较真实地反映簇的边界情况，因而适合于非球形簇的表示。基于密度的聚类算法大多采用核心点或单元表示簇。其中，NN 方法采用核心点表示簇；而基于单元的方法采用密集单元表示簇，但由于单元只是对落入其中的数据点的近似描述，因此只能近似反映簇的情况。

上述表示簇的方法都是结合具体的聚类算法提出的，不具有通用性。为了使聚类结果更易理解，可以采用可视化方法[4]，比如相同簇中的数据使用相同的标记显示，而不同簇中的数据使用不同的标记显示。

1.3 聚类分析面临的挑战

在数据挖掘领域，研究工作已经集中在为大型数据库有效和实际的聚类分析寻找适当的方法。活跃的研究主题集中在聚类方法的可伸缩性，方法对聚类复杂形状和类型的数据的有效性，高维聚类分析技术，以及针对大型数据库中混合数